QED: A Framework and Dataset for Explanations in Question Answering

Matthew Lamm[1]*, **Jennimaria Palomaki**[2], **Chris Alberti**[2],
Daniel Andor[2], **Eunsol Choi**[3]†, **Livio Baldini Soares**[2], **Michael Collins**[2]

[1]Department of Linguistics, Stanford University, United States
[2]Google Research, United States
[3]Department of Computer Science, The University of Texas at Austin, United States
{mrlamm, jpalomaki, chrisalberti, andor, liviobs, mjcollins}@google.com,
eunsol@cs.utexas.edu

Abstract

A question answering system that in addition to providing an answer provides an *explanation* of the reasoning that leads to that answer has potential advantages in terms of debuggability, extensibility, and trust. To this end, we propose QED, a linguistically informed, extensible framework for explanations in question answering. A QED explanation specifies the relationship between a question and answer according to formal semantic notions such as referential equality, sentencehood, and entailment. We describe and publicly release an expert-annotated dataset of QED explanations built upon a subset of the Google Natural Questions dataset, and report baseline models on two tasks—post-hoc explanation generation given an answer, and joint question answering and explanation generation. In the joint setting, a promising result suggests that training on a relatively small amount of QED data can improve question answering. In addition to describing the formal, language-theoretic motivations for the QED approach, we describe a large user study showing that the presence of QED explanations significantly improves the ability of untrained raters to spot errors made by a strong neural QA baseline.

1 Introduction

Question answering (QA) systems can enable efficient access to the vast amount of information that exists as text (Rajpurkar et al., 2016; Kwiatkowski et al., 2019; Clark et al., 2019; Reddy et al., 2019, among others). Modern neural systems have made tremendous progress in QA accuracy in recent years (Devlin et al., 2019). However, they generally give no explanation or justification of how they arrive at an answer to a question. Models that in addition to providing an answer can explain their reasoning may have significant benefits pertaining to trust and debuggability (Doshi-Velez and Kim, 2017; Ehsan et al., 2019).

Critical questions then, are what constitutes an *explanation* in question answering, and how we can enable models to provide such explanations. In an effort to make progress on these questions, in this paper we (1) introduce QED,[1] a linguistically grounded definition of explanations for extractive QA; and (2) describe an expert-annotated corpus of QED annotations based on the Natural Questions (Kwiatkowski et al., 2019) dataset. The QED corpus has been released publicly.[2]

Figure 1 shows a QED example. Given a question and a passage, QED represents an explanation as a combination of discrete, human-interpretable steps: (1) identification of a sentence implying an answer to the question, (2) identification of noun phrases in both the question and answering sentence that refer to the same thing, and (3) confirmation that the predicate in the sentence entails the predicate in the question once referential equalities are abstracted away.

This choice of explanation makes use of core semantic relations—referential equality and entailment—and thus has well-understood formal properties. In addition, we found that this way of decomposing explanations has high coverage

*Work done during internship at Google.
†Work done at Google.

[1]QED stands for the Latin ‘‘quod erat demonstrandum’’ or ‘‘that which was to be shown’’.

[2]https://github.com/google-research-datasets/QED.

Transactions of the Association for Computational Linguistics, vol. 9, pp. 790–806, 2021. https://doi.org/10.1162/tacl_a_00398
Action Editor: Preslav Nakov. Submission batch: 3/2021; Revision batch: 3/2021; Published 8/2021.

Question: who wrote the film howl's moving castle?
Passage: Howl's Moving Castle is a 2004 Japanese animated fantasy film written and directed by Hayao Miyazaki. It is based on the novel of the same name, which was written by Diana Wynne Jones. The film was produced by Toshio Suzuki.
Answer: Hayao Miyazaki

(1) Sentence Selection
Howl's Moving Castle is a 2004 Japanese animated fantasy film written and directed by Hayao Miyazaki.
(2) Referential Equality
the film howl's moving castle = Howl's Moving Castle
(3) Entailment
X is a 2004 Japanese animated fantasy film written and directed by ANSWER.⊢ ANSWER wrote X.

Figure 1: QED explanations decompose the question-passage relationship in terms of referential equality and predicate entailment.

(77% on the Natural Questions corpus[3]) and can be readily extended to other forms of question answering. (See Section 6.) Since QED decomposes the QA process into distinct subproblems, we also believe that it should enable research directions aimed at extending or improving upon extant QA systems.

In what follows, we present a definition of QED explanations. We then describe the dataset of QED annotations (7638/1353 train/dev examples), including discussion of the distribution of linguistic phenomena exhibited in the data. We move to propose four potential tasks, of varying complexity, related to the QED framework, and use the QED annotations to train and evaluate different models on two of these. Additionally, we describe a rater study which shows how the presence of QED explanations can help users identify errors made by an automated QA system.

2 Annotation Definition

We now describe the form of QED annotations. The treatment in this section is somewhat informal; for formal definitions see Appendix A.

2.1 Basic Definitions

We will use the following example to illustrate the approach:

[3]Instances with annotated short answers, omitting table passages.

Question: how many seats in university of michigan stadium
Passage: Michigan Stadium, nicknamed ''The Big House'', is the football stadium for the University of Michigan in Ann Arbor, Michigan. It is the largest stadium in the United States and the second largest stadium in the world. Its official capacity is 107,601.

The annotator is presented with a question/passage pair. Annotation then proceeds in the following four steps:

(1) Single Sentence Selection. The annotator identifies a single sentence in the passage that entails an answer to the question assuming that coreference and bridging anaphora (see later in this section) have been resolved in the sentence.[4]

In the above example, the following sentence entails an answer to the question, and would be selected by the annotator:

Its official capacity is 107,601.

This follows because given the passage context, ''Its'' refers to the same thing as the NP ''university of michigan stadium'' in the question, and the predicate in the sentence, ''X's official capacity is 107,601'', entails the predicate in the question ''how many seats in X''.

(2) Answer Selection. The annotator highlights a short answer span (or spans) in the answer sentence. In the above example the annotator would mark the following (answer shown with *[=A . . .]*):

Its official capacity is [=A 107,601].

(3) Identification of Question–Sentence Noun Phrase Equalities. The annotator marks referentially equivalent noun phrases, or noun phrases that refer to the same thing, in the question and the answer sentence. This includes reference not only to individuals and other proper nouns, but also to generic concepts.

In our example the annotator would mark the following two noun phrases (marked with the *[=1 . . .]* annotations) as referentially equivalent:

[4]If it is not possible to find a sentence that satisfies these properties—typically because the answer requires inference beyond coreference/bridging that involves multiple sentences—the annotator marks the example as not possible. See Section 3.

how many seats in [=1 university of michigan stadium]

[=1 Its] official capacity is [=A 107,601].

(4) Extraction of an Entailment Pattern. As a final, automatic step, an entailment pattern can be extracted from the annotated example by abstracting over referentially equivalent noun phrases. In the above example the entailment pattern would be as follows:

how many seats in X
X's official capacity is [=A 107,601].

2.2 Two Extensions

There are two extensions to the above approach, *coreference in answers*, and *bridging in referential equalities*:

Coreference in Answers. Consider the following example:

Question: who won wimbledon in 2019
Passage: Simona Halep is a female tennis player. She won Wimbledon in 2019.

In this case the single sentence *She won Wimbledon in 2019* would be selected by the annotator in step 1, as once coreference is resolved, this entails the answer to the question. The QED annotation would be as follows:

who won [=1 wimbledon] in [=2 2019]

[=A She (=C Simona Halep)] won [=1 Wimbledon] in [=2 2019]

In this case the answer ''She''—the substring in the original sentence—is not sufficient, as it involves an unresolved anaphor. Because of this, the annotator would mark the fact that ''She'' refers to ''Simona Halep'' earlier in the passage, using the *(=C . . .)* notation.

Bridging in Referential Equalities. Bridging anaphora (Clark, 1975) are frequently encountered in the QA passages in our data, and in Wikipedia more broadly. Consider the following:

Question: who won america's got talent season 11
Passage: The 11th season of America's Got Talent, an American talent show competition, began broadcasting in the United States during 2016. Grace VanderWaal was announced as the winner on September 14, 2016.

It is clear from the context surrounding the sentence ''Grace VanderWaal was announced as the winner on September 14, 2016'' that the noun phrase ''the winner'' refers to ''the winner of America's Got Talent Season 11'', and hence the sentence provides an answer to the question. It is helpful to imagine that there is an implicit prepositional phrase ''of America's Got Talent Season 11'' modifying ''the winner''. In this case the annotation would be the following:

who won [=1 america's got talent season 11]

[=A Grace VanderWaal] was announced as [=B the winner (of =1)] on September 14, 2016.

Here the annotation *[=B the winner (of =1)]* indicates that the phrase marked *[=1 . . .]* in the query is a bridged modifier to the phrase *the winner*, through the preposition ''of''.

Sometimes, there is no phrase like ''the winner'' above, but the referent is clearly an implicit argument of the supporting sentence. In this case we treat it as a bridge into the entire sentence.

2.3 A Note on Terminology

In defining QED we use the terms ''predicate'' and ''entailment'' in ways that may seem unfamiliar, but are not unrelated to the typical senses of those terms in linguistics. Canonically speaking, one thinks of a predicate as the semantic correlate of a verb in a sentence, and usually containing information about its argument structure. By taking a less structured notion of the term, as everything in a sentence surrounding a set of salient referring expressions, we are able to strike a balance between completely unstructured text, and more elaborate, structured representations that tend to be brittle.

The sense of entailment we intend then follows from this definition. A sentence entails an answer to a question if, having resolved and abstracted away referential equalities between the two, one can identify an answer to the question in the sentence.

3 QED Annotations for the Natural Questions

We now describe QED annotations over the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019). We first describe the annotation process,

then describe agreement statistics and statistics of types of referential expressions. For discussion of the assumptions we make and future extensions to QED, please see Section 6.3.

We focus on questions in the NQ corpus that have both a passage and short answer marked by the NQ annotator. We exclude examples where the passage is a table. A QED annotator was presented with a question/paragraph pair. Before performing the core QED annotation, annotators first determine whether: (1) there is a valid short answer within the paragraph (note that they can overrule the original NQ judgment), and there is a valid QED explanation for that answer; (2) there is a valid short answer within the paragraph, but there is no valid QED explanation for that answer; or (3) there is no valid short answer within the passage (hence the original NQ annotation is judged to be an error). Ten percent of all examples fell into category (3). Of the remaining 90% of examples that contained a correct short answer, 77% fell into category (1), and 23% fell into category (2).

Three QED annotators[5] annotated 7638 training examples (5154/1702/782 in categories 1/2/3 respectively), and 1353 dev examples (1019/183/151 in categories 1/2/3), without replication. We estimate that annotators averaged approximately 2 minutes per instance. Additionally, early stages of annotation consisted of regular adjudication among annotators to establish a consensus on QED's guidelines.

3.1 Agreement Statistics

All three annotators marked a common set of 100 examples drawn from the development set. We compute average pairwise agreement by comparing each annotator against the other two, and averaging across all pairs. Average classification of instances was 73.9%. If this seems low, it is because one annotator was more conservative interpreting QED's single sentence assumption, and pairwise accuracy breakdown was thus 81.2/72.3/68.1%. Given the high number of ''debatable'' instances reported in the Natural Questions paper, this divergence is unsurprising, however. Average pairwise F1 on mention identification/mention alignment, conditioned on both annotators labeling instances as amenable to QED, was 88.4 and 84.1, respectively.

[5]Three of the authors of this paper.

3.2 Types of Referential Expressions

The referential equality annotations are a major component of QED. Figure 2 shows some full QED examples from the corpus, and Figure 3 shows some example referential equalities from the corpus. In this section, in an effort to gain insight about the types of phenomena present, we describe statistics on types of referential equalities. We subcategorize referring expressions into the following types:[6]

Proper Names Examples are ''How I met your Mother'' or ''the cbs television sitcom how i met your mother''.

Non-Anaphoric Definite NPs These are expressions such as ''the president of the United States'' or ''the next Maze Runner film''. The majority involve one or more common nouns (e.g., ''president'', ''film'') together with a proper name, thereby defining a new entity that is in some sense a ''derivative'' of the underlying proper name.

Anaphoric Definite NPs These are definite NPs, most often from within the passage rather than the question, that require context to be interpreted. Examples are ''the series'' referring to an earlier mention of ''the Vampire Diaries'' within the passage, or ''the winner'' referring to ''the winner of America's got Talent Season 11''.

Generics Examples are ''a dead zone'' in the question ''what causes a dead zone in the ocean'', or ''Dead zones'' in the passage sentence ''Dead zones are low-oxygen areas caused by . . . ''.

Pronouns Examples are it, they, he, she.

Bridging Referential expressions in the passage sentence that use bridging (see Section 2.2).

Miscellaneous All referential expressions not included in the categories above.

Table 1 shows the frequency distribution of per-instance referential equality counts. Figure 4 shows an analysis of 100 referential equality annotations from QED, with a breakdown by type of referring expression in the question and passage. Proper names, non-anaphoric definites,

[6]For formal discussion, see Carlson (1977), Krifka (2003), Abbott (2004), and Mikkelsen (2011), among others.

Pronominal reference

Question: how many blocks in the great pyramid of giza$_1$

Wikipedia page: Great_Pyramid_of_Giza

Passage: Based on these estimates, building the pyramid in 20 years would involve installing approximately 800 tonnes of stone every day. Additionally, since **it**$_1$ consists of an estimated 2.3 million$_A$ blocks, completing the building in 20 years would involve moving an average of more than 12 of the blocks into place each hour, day and night. The first precision measurements of the pyramid were made by Egyptologist Sir Flinders Petrie in 1880–82 and published as The Pyramids and Temples of Gizeh. Almost all reports are based on his measurements[...]

Inexact match

Question: where does **the term sixes and sevens**$_1$ originate

Wikipedia page: At_sixes_and_sevens

Passage: An ancient dispute between the Merchant Taylors and Skinners livery companies$_A$ is the probable origin of **the phrase**$_1$. The two trade associations, both founded in the same year (1327), argued over sixth place in the order of precedence. In 1484, after more than a century and a half of bickering, the Lord Mayor of London Sir Robert Billesden ruled that at the feast of Corpus Christi, the companies would swap between sixth and seventh place and feast in each other's halls[...]

Answer bridging/coreference

Question: what is whitney houston's mother$_1$'s name

Wikipedia page: Cissy_Houston

Passage: **Emily "Cissy" Houston**$_A$ (née Drinkard; born September 30, 1933) is an American soul and gospel singer. After a successful career singing backup for such artists as Dionne Warwick, Elvis Presley and Aretha Franklin, Houston embarked on a solo career, winning two Grammy Awards for her work. **Houston** is the mother of singer Whitney Houston$_1$, grandmother of Whitney's daughter, Bobbi Kristina Brown, aunt of singers Dionne and Dee Dee Warwick, and a cousin of opera singer Leontyne Price.

Entity bridging

Question: who sang the national anthem$_1$ at **the first game of 2017 world series**$_2$

Wikipedia page: 2017_World_Series

Passage: **Game 1**: The ceremonial first pitch was thrown out by members of former Dodger Jackie Robinson's family, including his widow Rachel. The game marked the 45th anniversary of Robinson's death, and the 2017 season was the 70th anniversary of his breaking of the baseball color line. **[(at . . .)]**$_2$ Keith Williams Jr.$_A$, a gospel singer, performed "The Star-Spangled Banner", the national anthem$_1$.

Generic reference

Question: what is the function of **a paints binder**$_1$

Wikipedia page: Paint

Passage: **The binder**$_1$ is the film-forming$_A$ component of paint. It is the only component that is always present among all the various types of formulations. Many binders are too thick to be applied and must be thinned. The type of thinner, if present, varies with the binder.

Figure 2: Examples from the QED dataset, grouped according to different types of referential equalities.

and generics dominate expression types in the question (73, 16, and 6 examples, respectively). Expressions in the sentence are more diverse, with a much greater proportion of anaphoric definites, pronouns, and bridging examples (21, 9, and 5 cases, respectively). Finally, as an indication of the difficulty of the referential equality task, we note that in only 12% of all referential equalities in the 100 examples in Figure 4 is there an exact string match (after lower-casing of both question and passage) between the question and passage referential expression.

4 Tasks and Models

The QED data, which we release publicly, can be used as part of wide range of QA tasks and models. After discussing some of these tasks, we assess how well two recent neural architectures, one structured and one sequence-to-sequence, perform on two of them.

4.1 QED-based Tasks

Each QED example is a (q, d, c, a, e) tuple where q is a question from the NQ dataset, d is a Wikipedia

Question Expression	Passage Expression
how i.met your mother	the CBS television sitcom How I Met Your Mother
the most wins in the nfl	most wins
mantis	Mantis
the nashville sound	Countrypolitan - a smoother sound typified through the use of lush string arrangements with a real orchestra and often, background vocals provided by a choir
a permit driver	a driver operating with a learner 's permit
god's not dead a light in the darkness	it
the current president of un general assembly	the United Nations General Assembly President of its 72nd session beginning in September 2017
the new maze runner movie	Runner : The Death Cure
a box lacrosse team	a team

Figure 3: Referential equalities from the QED corpus.

page, c is a passage within d,[7] a is a short answer within c, and e is a QED explanation. A formal definition of e can be found in Appendix A. In brief, it consists of a sentence, which is a span within c, as well as a set of referential equalities. Each referential equality is a pair consisting of a question span together with a passage span (or a bridging position in the passage). Additionally, where an answer span a falls outside of the selected sentence, the explanation contains an answer coreference span.

We use $\mathcal{E}$ to refer to set of evaluation examples (either the development or test set). We focus our modeling efforts on the following two tasks, in order of increasing complexity:

Task 1: Explanation Prediction given Short Answer Given a (q, d, c, a) 4-tuple, make a prediction $\hat{e} = f(q, d, c, a)$ where f is a function that maps a (q, d, c, a) tuple to an explanation. We might, for example, define $f(q, d, c, a) = \arg\max_e p(e|q, d, c, a; \theta)$ under some model $p(\ldots)$. The evaluation measure is then $\sum_{(q,d,c,a,e)\in\mathcal{E}} l_1(e, f(q, d, c, a))$ where $l_1(e, \hat{e})$ is a per-example evaluation measure indicating how close $\hat{e}$ is to e.

[7]Passages are the same as NQ long answers.

	Referential Link Count			
	0	1	2	3
Instances	54	649	294	6

Table 1: Referential link count frequency distribution in a random sample of 1000 instances. When there are 0 links, the explanation consists only of a selected sentence.

Qu. \ Ps.	P	N	A	G	Pn	B	M	T
Proper	44	0	16	0	9	4	0	73
Def. (Non-Ana)	4	6	4	0	0	1	1	16
Def. (Ana)	0	1	1	0	0	0	0	2
Generic	0	0	0	6	0	0	0	6
Pronoun	0	0	0	0	0	0	0	0
Bridge	0	0	0	0	0	0	0	0
Misc	2	0	0	0	0	0	1	3
Total	50	7	21	6	9	5	2	100

Figure 4: Counts for 100 randomly drawn referential equality annotations from the QED corpus, subcategorized by expression type in the question (Qu.) and passage (Ps.). P/N/A/G/Pn/B/M refer to Proper/Def (non-ana)/Def(ana)/Generic/Pronoun/Bridge/Misc.

Task 2: Joint Answer and Explanation Prediction Given a (q, d, c) triple, predict $(\hat{a}, \hat{e}) = f(q, d, c)$, where f is a function that maps it to a short-answer/explanation pair. We might, for example, define $f(q, d, c) = \arg\max_{a,e} p(a, e|q, d, c; \theta)$ under some model $p(\ldots)$. The evaluation measure is $\sum_{(q,d,c,a,e)\in\mathcal{E}} l_2((a, e), f(q, d, c))$ where l_2 is some per-example measure.

By extension, one can conceive of a task in which one must also predict a passage c, in addition to an answer and an explanation. One could even integrate QED with a version of the open-domain QA task, which also entails retrieval of documents d. Given QED's linguistic generality, the data may also be useful as auxiliary input for training models that are not explicitly interested in evaluating explanation generation.

An open question in explainability is how we can build and evaluate models that generate *faithful* explanations, where the explanation truly reflects the model's underlying reasoning (Jacovi and Goldberg, 2020). Accurate models for the above tasks, even if they do not generate faithful explanations, may still have considerable utility. However, faithful models have several desirable

characteristics (see Sections 5 and 6); we view them as a major avenue for future work.

4.2 A SpanBERT Model

The first model we consider for Tasks 1 and 2 uses the SpanBERT coreference resolution model (Joshi et al., 2020; Lee et al., 2017) to identify referential equalities, extends the model with a QA component, and heuristically selects supporting sentences to produce a final QA+QED output.

Representation Assume an example contains a question q of m tokens $q_1 \ldots q_m$ and a passage c consisting of n tokens $c_1 \ldots c_n$. We denote the title of the Wikipedia page separately as the sequence t of k tokens $t_1 \ldots t_k$. The model uses SpanBERT to jointly encode the the concatenation of these token sequences,

$$\texttt{[CLS]}\, t_1 \ldots t_k \,\texttt{[S1]}\, q_1 \ldots q_m \,\texttt{[S2]}\, c_1 \ldots c_n \,\texttt{[SEP]}$$

as an input document.[8]

Coreference Given some document d and a candidate mention x, corresponding to a span within d, define $\mathcal{Y}(x)$ to be the set of potential antecedents for x. Each antecedent is either a span in the document with start-point before x in the document, or ϵ signifying that x does not have an antecedent. We can then define a distribution over the antecedent spans $\mathcal{Y}(x)$ as $p(y|x, D) = \frac{e^{s(x,y)}}{\sum_{y' \in \mathcal{Y}(x)} e^{s(x,y')}}$ where

$$s(x,y) = \begin{cases} 0 & \text{if } y = \epsilon; \\ s_m(x) + s_m(y) + s_c(x,y) & \text{otherwise} \end{cases}$$
$$s_m(x) = \texttt{FFNN}_m(g_x)$$
$$s_c(x,y) = \texttt{NN}_c(g_x, g_y)$$

where g_x and g_y are span representations obtained by concatenating the SpanBERT representations of the first and last token in each mention span. The scoring functions s_m and s_c represent mention and joint span match scores respectively. Whereas s_m is a simple feedforward net, s_c is a more complex scoring function that has been optimized to the coreference task. We refer the reader to Lee et al. (2017) for more details.

Lee et al. (2017) describe a method for training the model based on log-likelihood, and a beam search method that uses the scores $s_m(\ldots)$ and $s_c(\ldots)$ to filter candidate mention, antecedent pairs into the final set considered by the loss function. The final output from the coreference model is a hard clustering of the potential mentions into coreference clusters.

Given the constraints of QED referential equalities, we restrict s_c to only score coreferential links between the query and the passage or between the query and the title (all other values for s_m or s_c are set to $-\infty$). We model bridges as links between a query passage the title.

We finally post-process the cluster outputs of the coreference component as follows: For each cluster we output the first mention in the cluster that appears in the question with the first mention in the cluster of references that appears in the passage, once cluster mentions are sorted.[9] If there is no cluster mention in the passage, we assume the passage reference is a bridge.

QA The answer scoring component computes answer candidate representations g_z using the same candidate mention scoring network as the coreference model, $\texttt{FFNN}_m$, as well as a feed-forward network, $\texttt{FFNN}_q$, that scores candidate answer spans relative to a representation of the question. The score of an answer z is then computed as

$$s_a(z) = \texttt{FFNN}_m(g_z) + \texttt{FFNN}_a(g_z, g_q).$$

Thus, the only new parameters belong to a single hidden layer feed-forward net $\texttt{FFNN}_a$ that specifically targets the question-answer relationship. Apart from the use of shared candidate mention scoring parameters, no further dependence is introduced between the answer and referential equality predictions.

Sentence Selection We perform sentence selection heuristically by choosing the sentence containing the first cluster output by the coreference model. Any subsequent coreference cluster containing a document mention outside of this sentence is dropped in the final prediction. If no referential link is predicted, we take the supporting sentence to be the one containing the answer span.

Training For Task 1, we consider an untrained model and a fine-tuned model, both of which omit the QA component described above. In the former, we do not use expert annotated QED data but instead use the CoNLL OntoNotes coreference

[8] We use [S1] = ‘‘.’’ and [S2] = ‘‘?’’ as separators.

[9] This is necessary because it is technically possible for a cluster to contain more than two mentions before post-processing.

dataset (Pradhan et al., 2012) to train the pretrained SpanBERT model. We only score document mentions in the sentence containing the answer.

For the fine-tuned model, we mark short answers with special tokens before computing the SpanBERT document representation. Then, we further train the model with the training portion of QED data. We used SpanBERT ''large'', with a maximum span width of 16 tokens, a top span ratio of 0.2, 30 max antecedents per mention. In fine-tuning, we used an initial learning rate of $3 \cdot 10^{-4}$ and trained for 3 epochs on the QED training set.

For Task 2, we train the QA and Coreference components in a multitask fashion, by minimizing the weighted sum of the QA and coreference cross entropy losses. For the QA data, we augment using passages containing short answers from NQ. Our best results are obtained with a weight of 5 on the coreference loss and 2 epochs of training. The best answer accuracy and QED F1 are obtained for different base learning rates of $2 \cdot 10^{-5}$ and $5 \cdot 10^{-5}$ respectively.

4.3 A T5 Model

The second model we consider fine-tunes T5 (Raffel et al., 2020) to predict *linearized* QA and QED outputs from an input document. We briefly describe the linearization approach here, and refer the reader to Appendix B for a worked example.

Input Representation Similar to the SpanBERT model and as depicted immediately above, we pass the concatenation of question, title, and document tokens as input to T5, in that order.[10] Each input instance is either a QA- or QED-specific instance, which is indicated to T5 by appending a task-specific token to the end of the input.

Output Representation The model is tasked with predicting either (1) an answer span or (2) a QED explanation, represented as a sequence of referential equalities, all separated by a special token.[11] In (2), each referential equality is represented as the concatenation of two spans: the tokens in its query mention and the tokens in its passage mention, separated by ">>". In both (1) and (2) the four tokens in the passage immediately following the answer or passage mention are also appended. These additional tokens are not part of the evaluated spans; they serve to uniquely locate the character offset of the answer or passage mention during evaluation.

Sentence selection proceeds heuristically as in the SpanBERT model.

[10]We use ">>" as field separators.

[11]We use "&&" to separate referential equalities.

	Mention Identification			Mention Alignment			Sentence Accuracy
	P	R	F1	P	R	F1	
SB-onto	59.0	35.6	44.4	47.7	28.8	35.9	97.3
SB-fine-tuned	76.8	68.8	72.6	68.4	61.3	64.6	94.2
T5	73.0	75.8	74.4	63.1	65.5	64.3	95.9

Table 2: SpanBERT (SB) and T5 11B model performance for Task 1: recovering QED annotations when the correct answer is given.

Training We trained T5 11B on only the QED training data, using the standard fine-tuning recipe with a batch size of 1024, learning rate of $2e{-}4$ and a dropout rate of 0.1. For Task 1, we trained on the explanation task, marking short answers using "«" and "»" brackets in the input. For Task 2, we mixed the QA task and the explanation task with equal weights, and randomly shuffled the instances. We saw the best results when we trained Task 1 for 7000 steps and Task 2 for 2000 steps.

4.4 Evaluation and Results

We evaluate answer selection, sentence selection, and the identification of referential equalities. For answer and sentence selection, we report accuracy on 90% span overlap F1. For referential equality, we evaluate both mention identification (the identification of individual referential expressions in the question and passage) and referential equality detection (the identification of pairs of referential expressions).[12] We compute precision, recall, and F1 measure in both cases.[13]

Results for Task 1 for both the SpanBERT and T5 models are reported in Table 2. The table shows results for both the OntoNotes- and QED-fine-tuned SpanBERT models, as well as the T5 model trained only on the task of explanation prediction. Of note is that trained models trained on QED data do considerably better than the model trained on OntoNotes, indicating that referential

[12]Where referential equalities involved bridged passage mentions, we only evaluate the models' ability to recognize that they are bridged, since there are many conceivable places in a sentence into which mentions can be bridged.

[13]Official evaluation code has been released with the dataset.

	Mention Identification			Mention Alignment			Answer Accuracy	Sentence Accuracy
	P	R	F1	P	R	F1		
SB-QED-only	74.0	63.1	68.1	63.6	54.2	58.6	–	88.4
SB-QA-only	–	–	–	–	–	–	73.0	81.5
SB-QA+QED	77.6	64.4	70.4	68.9	57.2	62.5	74.5	90.8
T5-QED-only	71.1	73.3	72.2	59.5	61.4	60.4	–	88.9
T5-QA-only	–	–	–	–	–	–	78.9	88.7
T5-QA+QED	70.3	72.3	71.3	58.3	59.9	59.1	79.2	89.1

Table 3: SpanBERT (SB) and T5 model performance for Task 2: recovering answer and QED annotations given a passage that is known to contain the answer.

equalities are of a distinct distribution from other coreference data.

In Table 3 we report results for Task 2. SB-QED-only refers to the SpanBERT model fine-tuned only with QED data. SB-QA-only refers to the SpanBERT model fine-tuned on the NQ QA data. SB-QA+QED finetunes on both QA and QED. Similarly, we report results for T5 models. We find that the T5 model tends to have higher recall than the SpanBERT model on mention evaluations, but that the SpanBERT model is considerably more precise. T5 far outperforms SpanBERT on answer accuracy, even though it was fine-tuned without the NQ QA data. Interestingly, in both T5 and SpanBERT models, training on QED data improves QA performance. While the SpanBERT model is more complex than the sequence-to-sequence T5 model, it is considerably more compact (320 million parameters versus 11 billion).

The data contain annotations for answer coreference, in which answer spans outside of the supporting sentence are referred to by an anaphor within it. (See Section 2.) The phenomenon is relatively rare though, and hence there are not enough data to evaluate performance properly. We did perform an additional experiment with T5 where, in addition to an answer span, it predicted its anaphor in the answering sentence where appropriate. The model achieved satisfactory performance, with an F1 of 71%.

5 Rater Study

A major desideratum for explanation generation models is faithfulness—that is, when the explanations generated by a model truly reflect its reasoning process (Jacovi and Goldberg, 2020; Ross et al., 2017). One motivation for this is that when a model is wrong, faithful explanations reliably indicate the reason for the error. In the context of QA, exposing the explanations of a faithful system should improve users' ability to spot incorrect answers. We show that this is true of a faithful QED system using a rater study.

5.1 Task Setup

Given a question, passage, and a candidate answer span, raters were tasked with assessing whether the candidate answer was correct or incorrect, and indicating the confidence of their assessment.

A total of 354 raters, all of whom are US residents and native English speakers, were divided into three disjoint pools to perform the task in three distinct test settings: The **None** group of raters ($n = 121$) was presented with a question, passage, and a highlighted answer span. The **Sentence** group ($n = 117$) was provided with additional highlighting of the sentence justifying the answer, with no distinction made between referential equalities and predicates. The **QED** group ($n = 116$) was provided with additional highlighting to indicate referential equalities between spans in the question and spans in the passage. On average, a given rater provided judgments for 41 questions.

We constructed the data for the study by taking a random set of 50 correct answers, and 50 incorrect guesses from the NQ baseline model (Alberti et al., 2019), on the Natural Questions dev set. So as to ensure that the task was sufficiently challenging, correct instances were the gold answer spans on question/passage pairs where the model produced a false negative—that is, where an answer existed in the passage, but the model was not confident about it. Incorrect instances were false positive guesses from the model, where an answer did not exist in the passage but the model was confident that one did.

Explanations, where present, were manually annotated to simulate the inferences of a hypothetical model that used a QED-style reasoning process. When an item's answer was correct, the explanation shown was simply its corresponding QED explanation. When the answer was incorrect, referential equalities were identified using counterfactual reasoning; they indicate equivalences that *would have to hold* if the answer *were* correct.

Representative examples from the set of rater study items are shown in Figure 5. Note that although referential equalities were manually chosen for incorrect examples, they are not outlandish: They tend to correspond with closely

Correct answer

Question: where is nathan's hotdog eating contest$_1$ held

Passage: The Nathan's Hot Dog Eating Contest is an annual American hot dog competitive eating competition. It$_1$ is held each year on Independence Day at Nathan's Famous Corporation's original, and best-known restaurant at the corner of Surf and Stillwell Avenues in Coney Island$_A$, a neighborhood of Brooklyn, New York City.

Referential equality error

Question: whats on top of the white house$_1$

Passage: The Statue of Freedom$_A$, also known as Armed Freedom or simply Freedom, is a bronze statue designed by Thomas Crawford (1814 – 1857) that, since 1863, has crowned the dome of the U.S. Capitol building in Washington , D.C.$_1$ Originally named Freedom Triumphant in War and Peace, a U.S. government publication now states that the statue " is officially known as the Statue of Freedom ". The statue depicts a female figure wearing a military helmet and holding a sheathed sword in her right hand and a laurel wreath and shield in her left.

Predicate entailment error

Question: what percentage of the us population$_1$ is over 50

Passage: There were about 125.9 million adult women in the United States in 2014. The number of men was 119.4 million. At age 85 and older, there were almost twice as many women as men (4 million vs. 2.1 million). People under 21 years of age made up over a quarter of the U.S. population$_1$ (27.1 %), and people age 65 and over made up one - seventh (14.5 %$_A$). The national median age was 37.8 years in 2015.

Figure 5: Three example items from the rater study. In the **referential equality error** example, the answer is incorrect because the White House and the State Capitol Building are not the same. In the **predicate entailment error** example, the answer is incorrect because the sentence mentions the number of people over 65, whereas the question asks for the number of people over 50.

related referents, but that are not equivalent upon further inspection. More generally, incorrect answers in the rater study tend to be incorrect for very subtle reasons; this is a result of the aforementioned answer selection process.

All raters were told that highlighting was the output of ''an automated question answering system'' that was incorrect ''about half of the time.'' They were advised not to use external knowledge sources or web search to make their judgments. Raters who saw explanations were also told that the system made use of the highlighted explanation to produce its candidate answers.

5.2 Results

Average rater accuracies for each test setting are presented in Table 4. We see that, in aggregate, QED explanations improved accuracy on the task over and above the other test settings, and gave the most improvement on the identification of answers that were incorrect. These improvements translate to incorrect answers resulting from both predicate and reference model errors.

Somewhat surprisingly, highlighting just the sentence containing the answer improved accuracy more than including referential equality highlighting on instances that were correct. This may be because raters' propensity to mark

	Accuracy			F1
	All	Corr	Incorr/Pred/Ref	Incorr
None	67.5	90.4	44.3/43.9/44.7	57.6
Sentence	69.7	**92.4**	47.1/46.1/48.0	60.9
QED	**70.2**	90.6	**49.7/48.2/51.0**	**62.5**

Table 4: Rater study results. Corr and Incorr are accuracies of raters in each group on correct and incorrect instances respectively, with incorrect instances further broken into Pred(icate) and Ref(erence) model errors. F1 is on the task of identifying incorrect instances.

instances correct decreases as the complexity of explanations increases, from None (73.1%) to Sentence (72.6%) to QED (70.5%).

Also clear from Table 4 is that rater accuracy is much lower on incorrect instances. Even though raters were told that the answers presented were incorrect half of the time, they judged the answers to be correct roughly 71% of the time.[14]

Figure 6 provides another perspective on the disparity in judgments on correct/incorrect

[14]While this confirmation bias presents an interesting challenge for future work, it is not a shortcoming of our results: Raters were not trained to do well on the task, as we aimed to approximate how users interact with automated QA systems.

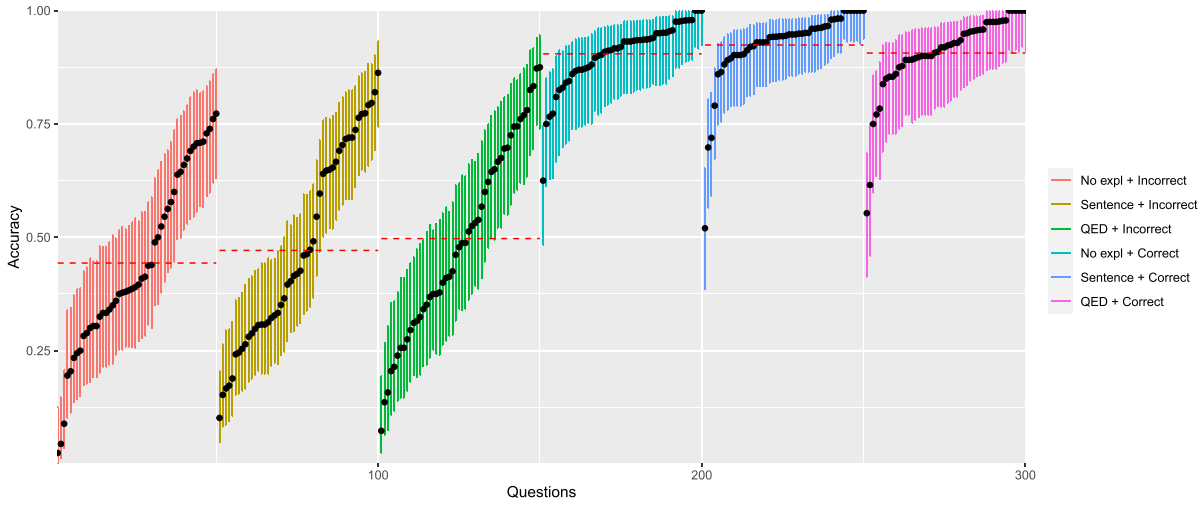

Figure 6: Sorted, per-question evaluation accuracies from different rater study settings, with 95% binomial confidence intervals. The ‘‘evaluation accuracy’’ for a question is the proportion of raters who judged it correctly. Left three plots correspond to trials with incorrect answers highlighted; right three plots to trials with correct answers highlighted. Dashed red lines correspond to the average accuracy for each setting, identical to the numbers in Table 4.

instances summarized in Table 4. Highest per-question accuracies in the incorrect pool were still lower than the average accuracy on all correct instances, and the lowest accuracy on incorrect instances is far lower than that of any of the correct instances. The wide distribution of accuracies on incorrect instances ($\sigma \approx 0.50$) seen in Figure 6 was also reflected in the rater pool ($\sigma \approx 0.45$). The challenging nature of incorrect instances speaks to the promise of improvements from QED explanations.

5.3 Effectiveness of explanations

How statistically significant are the results reported in Table 4? The 14,115 test instances were spread across 354 raters and 100 questions. We use the `rstanarm` R package (Goodrich et al., 2020) to fit a generalized linear mixed model (GLMM) that estimates the log-odds of rater accuracy on the basis of fixed effects (instance correctness and explanation type), while controlling for random effects of rater and question. (See Gelman and Hill (2006) for further discussion of GLMMs.) Ultimately we are interested in the magnitude and statistical properties of the model under the various test settings.

Table 5 shows the fixed effect coefficient and standard deviations for each setting. The presence of QED explanations in the Incorrect setting increased the log-odds of rater accuracy by 0.25, with a posterior predictive p-value of 0.015 that this effect is greater than zero. The comparable effect for Sentence explanations was 0.15, with a posterior predictive p-value of 0.08. The rater and question random effects had standard deviations of 0.63 and 0.90 respectively, reflecting again the high variance of questions shown in Figure 6.

Parameter	Coefficient (SD)
(Intercept)	−0.31 (0.15)
+ Incorrect+Sentence	0.15 (0.11)
+ Incorrect+QED	0.25 (0.11)
+ Correct+None	2.94 (0.21)
+ Correct+Sentence	3.04 (0.13)
+ Correct+QED	2.69 (0.13)

Table 5: Generalized linear mixed model fixed effect coefficients, showing mean and standard deviation of 10k MCMC samples. The Intercept corresponds to the Incorrect+None setting.

As we saw earlier, the effects of explanations in the Correct setting was reversed: The Sentence explanations caused a small, statistically insignificant increase in log-odds, while QED explanations caused a statistically significant drop in log-odds.

6 Discussion

6.1 QED versus other Explanation Types

QED exists in between relatively unstructured explanation forms on the one hand, such as attention

Multi-hop

Question: when did martial law$_1$ end in the philippines$_2$

Wikipedia page: Proclamation_No._1081

Passage: Proclamation No 1081$_3$ was the proclamation of martial law$_1$ in the Philippines$_2$ by President Ferdinand Marcos . [. . .]$_2$ It$_3$ was announced to the public on 23 September 1972 , and was formally lifted on 17 January 1981$_A$.

Yes/No

Question: can you make and receive calls in airplane mode$_1$

Wikipedia page: Airplane_mode

Passage: Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode$_1$ is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi.

Answer: NO

Set-valued

Question: who was involved in the soviet invasion of afghanistan$_1$

Wikipedia page: Soviet–Afghan War

Passage: The Soviet – Afghan War lasted over nine years , from December 1979 to February 1989 . [. . .]$_1$ Insurgent groups known as the mujahideen$_{A1}$ fought against the Soviet Army$_{A2}$ and the Democratic Republic of Afghanistan government$_{A3}$, mostly in the country 's rural countryside . [. . .]$_1$ The mujahideen groups were backed by the United States$_{A4}$ and Pakistan$_{A5}$, making it a Cold War proxy war .

Figure 7: Examples from NQ that go beyond the current definition of QED. Highlighting resembles QED highlighting. In the **multi-hop** instance, we require an additional sentence to link the entity mentions in the question to the answer sentence. In the **yes/no** question, the supporting sentence justifies a ''No'' answer because it contradicts the question predicate. In the **set-valued** question, multiple sentences provide partial answers to the question, and the resulting answer is the union of all of these.

distributions (Wiegreffe and Pinter, 2019; Jain and Wallace, 2019; Mohankumar et al., 2020) or sequential outputs (Camburu et al., 2018, 2020; Narang et al., 2020; Kumar and Talukdar, 2020) and more elaborate, discrete semantic representations that can in theory be applied to explainable QA (Abzianidze et al., 2017; Wolfson et al., 2020).

6.2 QED and Faithfulness

A major goal for future work is to develop faithful QA models with the QED framework. As Section 5 suggests, models that are not only right for the right reasons, but also wrong for the right reasons, can help users identify subtle errors. Other motivations include model debuggability: Since faithful models should reveal weaknesses in their reasoning, they may enable more targeted intervention.

QED is a promising style of explanation to this end, because it makes use of fundamental semantic variables, like reference (Russell, 1905; Clark and Marshall, 1981; Tomasello et al., 2007). We can say, definitively, that in order for a sentence to answer a question about a thing, its meaning must involve that thing in a very particular sense. Posed counterfactually, when you break referential equality, you break answerhood, and the same argument follows for predicate entailment. This is a hallmark of a good explanation (Pearl, 2019; Lipton, 2001).

6.3 Scoping and Extension to other Question Types

The instantiation of QED presented in the current work is limited to extractive wh-questions whose answers are entailed by single sentences. We feel this scoping is well justified, because (1) a significant portion of NQ falls under QED's current purview; (2) previous work and data analysis suggests QED can be readily extended to accommodate these other types (Hearst, 1992; Miltsakaki et al., 2004; Lamm et al., 2018; Tandon et al., 2019); and (3) close study of the single sentence case is a necessary condition for these other question types.

In Figure 7, we present several representative NQ instances that require more machinery than QED provides at present. Let us consider how QED might be extended to handle each of these.

Multi-hop QA For multi-hop questions, referential equalities involve longer, text-mediated paths from entity references in the question to an ultimate sentence entailing its answer (Yang et al., 2018).

Yes/No QA Answering Yes/No questions requires identifying sentences in a text that entail *or contradict* the premise presented in the question.

Set-valued QA Set-valued QA requires assembling QED explanations for *a set* of answers to the question, and returning the union of the (unique) answers found.

Looking further afield from these question types above, which are less frequent in NQ but nevertheless attested there, it becomes clear that QA writ large is much broader than even a dataset of NQ's scale suggests (Rogers et al., 2020). The generality of QED as a model for how elements of questions can link up with textual evidence suggests that QED would likely be complementary to, rather than at odds with, efforts to understand these broader senses of QA.

7 Conclusions

We have described QED, a framework for explanations in question answering, and we have introduced a dataset of QED annotations. The framework is grounded in referential equality, and entailment. In addition we have described baseline models for two QED-based tasks, and a rater study utilizing QED annotations.

Future work should consider the development of models based on QED, especially those that provide faithful explanations, and extensions of QED beyond the single-sentence assumption.

References

Barbara Abbott. 2004. Definiteness and Indefiniteness. *The Handbook of Pragmatics*, 122. https://doi.org/10.1002/9780470756959.ch6

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-2039

Chris Alberti, Kenton Lee, and Michael Collins. 2019. A BERT baseline for the natural questions. *arXiv preprint arXiv:1901.08634*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, pages 9539–9549.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! Adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.382

Greg N. Carlson. 1977. A unified analysis of the English bare plural. *Linguistics and Philosophy*, 1(3):413–457. https://doi.org/10.1007/BF00353456

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Herbert H. Clark. 1975. Bridging. In *Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing*, TINLAP '75, page 169–174, USA. Association for Computational Linguistics. https://doi.org/10.3115/980190.980237

Herbert H. Clark and Catherine R. Marshall. 1981. Definite reference and mutual knowledge. *Elements of Discourse Understanding*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated rationale generation: A technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 263–274. ACM. https://doi.org/10.1145/3301275.3302316

Andrew Gelman and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. 2020. rstanarm: Bayesian Applied Regression Modeling via Stan. R package version 2.19.3.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*. https://doi.org/10.3115/992133.992154

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.386

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77. https://doi.org/10.1162/tacl_a_00300

Manfred Krifka. 2003. Bare NPS: Kind-referring, indefinites, both, or neither? *Semantics and Linguistic Theory*, 13:180–203.

Sawan Kumar and Partha Talukdar. 2020. Nile: Natural language inference with faithful natural language explanations. https://doi.org/10.3765/salt.v13i0.2880

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466. https://doi.org/10.1162/tacl_a_00276

Matthew Lamm, Arun Chaganty, Christopher D. Manning, Dan Jurafsky, and Percy Liang. 2018. Textual analogy parsing: What's shared and what's compared among analogous facts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 82–92, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1008

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end ution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language*

Processing, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Peter Lipton. 2001, What good is an explanation?, Giora Hon and Sam S. Rakover, editors, *Explanation: Theoretical Approaches and Applications*, Springer Netherlands, Dordrecht, pages 43–59. https://doi.org/10.1007/978-94-015-9731-9_2

Line Mikkelsen. 2011. Copular clauses. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*, 2, pages 1805–1829, Berlin. Mouton De Gruyter.

Eleni Miltsakaki, Rashmi Prasad, Aravind K. Joshi, and Bonnie L. Webber. 2004. The Penn Discourse Treebank. In *LREC*.

Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2020. Towards transparent and explainable attention models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4206–4216, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.387

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! Training text-to-text models to explain their predictions.

Judea Pearl. 2019. The limitations of opaque learning machines. *Possible Minds: Twenty-Five Ways of Looking at AI*13–19.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes. In *EMNLP-CoNLL Shared Task*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/D16-1264

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266. https://doi.org/10.1162/tacl_a_00266

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8722–8731. https://doi.org/10.1609/aaai.v34i05.6398

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2662–2670.

Bertrand Russell. 1905. On denoting. *Mind*, 14(56):479–493. https://doi.org/10.1093/mind/XIV.4.479

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for ‘‘What if...’’ reasoning over procedural text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6078–6087. https://doi.org/10.18653/v1/D19-1629

Michael Tomasello, Malinda Carpenter, and Ulf Liszkowski. 2007. A new look at infant pointing. *Child Development*, 78(3):705–722. https://doi.org/10.1111/j.1467-8624.2007.01025.x, PubMed: 17516997

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1002

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198. https://doi.org/10.1162/tacl_a_00309

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.18653/v1/D18-1259

A A Formal Definition of QED Annotations

An annotator is presented with a question q that consists of m tokens $q_1 \dots q_m$, along with a passage c consisting of n tokens $c_1 \dots c_n$.

The QED annotation is a triple $\langle s, e, a \rangle$ where:

- s is a sentence within the passage c. Specifically s is a pair s_0, s_1 indicating that the sentence spans words $c_{s_0} \dots c_{s_1}$ inclusive.
- e is a sequence of 0 or more ''referential equality annotations'', $e_1 \dots e_{|e|}$. Each member of e specifies that some noun phrase within the question refers to the same item in the world as some noun phrase within the sentence s.
- a is one or more answer annotations $a_1 \dots a_{|a|}$.

We now describe the form of the e and a annotations, making reference to the following example. Subscripts indicate token positions:

Question: who$_1$ won$_2$ wimbledon$_3$ in$_4$ 2019$_5$
Passage: Simona$_1$ Halep$_2$ is$_3$ a$_4$ female$_5$ tennis$_6$ player$_7$.$_8$ She$_9$ won$_{10}$ Wimbledon$_{11}$ in$_{12}$ 2019$_{13}$.$_{14}$

As a preliminary step, given the paragraph c and sentence s, we use $\mathcal{S}$ to refer to the set of all phrases within s. Our initial definition of $\mathcal{S}$ is

$$\mathcal{S} = \{(i, j) : s_0 \leq i \leq j \leq s_1\}$$

We also define the set of question phrases $\mathcal{Q}$ and passage phrases $\mathcal{C}$ to be

$$\mathcal{Q} = \{(i, j) : 1 \leq i \leq j \leq m\}$$
$$\mathcal{C} = \{(i, j) : 1 \leq i \leq j \leq n\}$$

We can then give the following definitions:

Definition 1 *Each referential equality annotation e_k for $k = 1 \dots |e|$ is a pair $(\phi_k, \pi_k) \in \mathcal{Q} \times \mathcal{S}$, specifying that the phrase ϕ_k in the query refers to the same thing in the world as the phrase π_k within s.*

In our example,

$$e = [((3, 3), (11, 11)), ((5, 5), (13, 13))]$$

where the first tuple in the sequence corresponds with the alignment between ''wimbledon'' in the question and ''Wimbledon'' in the passage, and the second tuple with ''2019'' in the question and ''2019'' in the passage.

Definition 2 *Each answer annotation a_k for $k = 1 \dots |a|$ is a pair $(\pi_k, \xi_k) \in \mathcal{S} \times \mathcal{C}$ specifying that the answer is given by phrase π_k, and the full string corresponding to π_k after coreference is resolved is the phrase ξ_k. If no coreference resolution is required then $\pi_k = \xi_k$.*

In our example,

$$a = [((9, 9), (1, 2))]$$

corresponding with the alignment of ''She'' in the sentence ''She won Wimbledon in 2019'' with the mention of ''Simona Halep'' earlier in the passage.

A.1 Extending Annotations to Include Bridging

Recall the definition of bridging in Section 2. We extend the formal definition of QED to include bridging by redefining $\mathcal{S}$ to include implicit phrases introduced in the form of implicit prepositional phrases, as in the ‘‘winner [of ...]’’. The modified definition of $\mathcal{S}$ includes all phrases of the following form: (1) Any pair (i, j) such that $s_0 \leq i \leq j \leq s_1$ indicating the subsequence of words $c_i \dots c_j$ within the sentence. (2) Any triple (i, j, p) such that $s_0 \leq i \leq j \leq s_1$ and p is a preposition, indicating the implicit noun phrase in the sentence that modifies the phrase $c_i \dots c_j$ through the preposition p. (3) Any pair (NULL, p) such that p is a preposition, indicating the implicit noun phrase modifying the entire sentence $c_{s_0} \dots c_{s_1}$ through the preposition p.

Given the following example, then:

Question: who$_1$ won$_2$ america's$_3$ got$_4$ talent$_5$ season$_6$ 11$_7$
Passage: The$_1$ 11th$_2$ season$_3$ of$_4$ America's$_5$ Got$_6$ Talent$_7$ began$_8$ broadcasting$_9$ in$_{10}$ the$_{11}$ United$_{12}$ States$_{13}$ during$_{14}$ 2016$_{15}$.$_{16}$ Grace$_{17}$ VanderWaal$_{18}$ was$_{19}$ announced$_{20}$ as$_{21}$ the$_{22}$ winner$_{23}$ on$_{24}$ September$_{25}$ 14$_{26}$,$_{27}$ 2016$_{28}$.$_{29}$

we have that

$$e = [((3, 7), (22, 23), \text{‘‘of’’})]$$

This means that the question span ‘‘america's got talent season 11’’ is bridged by the reference ’’the winner’’ in the answering sentence. The preposition ‘‘of’’ indicates how the question referent can be attached to the sentence reference: Putting them together yields ‘‘the winner of america's got talent season 11’’.

B T5 Model Linearization

We describe our method for linearizing QED instances as T5 input and output sequences. Let us consider the following example:

Question: how many seats in university of michigan stadium
Passage: Michigan Stadium, nicknamed ‘‘The Big House’’, is the football stadium for the University of Michigan in Ann Arbor, Michigan. It is the largest stadium in the United States and the second largest stadium in the world. Its official capacity is 107,601.

Recall that the QED annotation for this example is as follows

how many seats in [=1 university of michigan stadium]

[=1 Its] official capacity is [=A 107,601]

The T5 model input are constructed by concatenating the question, page title, and paragraph into one sequence, separated by ‘‘>>’’:

how many seats in university of michigan stadium >> Michigan Stadium >> Michigan Stadium, nicknamed ‘‘The Big House’’, is the football stadium for the University of Michigan in Ann Arbor, Michigan. It is the largest stadium in the United States and the second largest stadium in the world. Its official capacity is 107,601.

A task-specific token is prepended to this input to indicate whether the model should produce an answer or an explanation. The answer output sequence is as follows:

107,601 ».

where additional material after the special character "»" (as distinct from ">>") is used to disambiguate the position of the answer in the passage. Finally, the explanation would be linearized as follows:

university of michigan stadium >> Its » official capacity is 107,601

Here, the first phrase corresponds to the question mention and the second to the passage mention. The additional material after the passage mention is meant to uniquely identify its position in the passage, for evaluation purposes.